

Mini-Minimax Uncertainty Quantification for Emulators

Jeffrey C. Regier[†] and Philip B. Stark[†]

Abstract. Consider approximating a function f by an *emulator* \hat{f} based on n observations of f . Let w be a point in the domain of f . The *potential error* of \hat{f} at w is the supremum of $|\hat{f}(w) - g(w)|$ among functions g that satisfy all constraints f is known to satisfy. The supremum over w of the potential error is the *maximum potential error* of \hat{f} . Suppose f is in a known class of regular functions. Consider the set \mathcal{F} of all functions in that class that agree with the n observations and are globally no less regular than those observations require f to be. We find a lower bound on the potential error of any emulator over $f \in \mathcal{F}$. This is the mini-minimax uncertainty. Its maximum over w lower-bounds the maximum potential error of any \hat{f} . If that is large, *every* emulator based on these observations is *potentially substantially incorrect*. To guarantee higher accuracy would require stronger assumptions about the regularity of f . We lower-bound the number of observations required to ensure that some emulator based on those observations approximates all $f \in \mathcal{F}$ to within ϵ . For the Community Atmosphere Model, the maximum potential error based on a particular set of 1154 observations of f is no smaller than the potential error based on a single observation of f at the centroid of the 21-dimensional parameter space. We also find lower confidence bounds for quantiles of the potential error and the mean potential error over w in the domain of f . For the Community Atmosphere Model, these lower confidence bounds are an appreciable fraction of the maximum potential error.

Key words. emulator, surrogate function, metamodel, minimax, Lipschitz, information-based complexity

AMS subject classifications. 68Q17, 65D05, 68U20, 62P12

1. Introduction. Emulators, also known as surrogate functions and metamodels, are important tools for approximating functions that have been observed only partially. Common emulation methods include Kriging, Multivariate Adaptive Regression Splines (MARS), Projection Pursuit Regression, Polynomial Chaos Expansions, as well as various Bayesian modeling techniques [1, 2, 3]. This paper studies the accuracy of emulators.

A common use for emulators is to approximate computer models. A computer model can be considered a deterministic function f from a vector of inputs in the domain of f to an output.¹ Only a finite number of simulations can be conducted, though typically an intractable number of inputs are possible—for instance if any input parameter is a floating point number. By fitting an emulator to observations of f , it becomes possible to cheaply approximate the computer model. While many computer models are effectively black boxes, emulators typically have closed forms, amenable to analytic study.

[†]Department of Statistics, University of California, Berkeley, CA 94720 (jeff@stat.berkeley.edu, stark@stat.berkeley.edu)

¹The function f might not be entirely deterministic, for instance, it could involve Monte Carlo simulations; moreover, in distributed parallel computations, numerical results can depend on the order in which subproblems happen to complete. These cases can be thought of as observing $f + \epsilon$, where ϵ is stochastic noise. We do not address the complication of noise here; however, uncertainty in the observations of f only makes the problem of approximating f to a given level of accuracy more difficult. Because we focus on lower bounds on the difficulty of approximating f accurately, our results generally remain lower bounds when the observations of f are not only incomplete, but also imperfect. It is only necessary to add the step of finding a lower confidence bound on the regularity of f from the observations, rather than finding the lower bound directly.

Computer models known as *HEB* [4] may be particularly difficult to emulate: They depend on *High-dimensional* inputs; they are *Expensive* to run; and they are effectively *Black boxes* that are not amenable to closed-form, analytic study. Because the models are high-dimensional, it takes prohibitively many runs to explore the domain of f —the number of runs needed grows exponentially in the dimension. Because the models are expensive, actually performing so many runs is impractical or impossible. And because the models are black boxes, there are few (if any) constraints that can be exploited to extrapolate reliably from inputs actually tried to inputs not sampled. There are numerous practical HEB problems, such as:

- Climate models: [5] (21–28-dimensional domains; 1154 simulations; Kriging and MARS)
- Automobile crashes: [6] (15-dimensional domain; 55 simulations; polynomial response surfaces and artificial neural networks).
- Chemical reactions: [7] (30–50-dimensional domain; boosted surrogate models) and [8] (46-dimensional domain; seconds per simulation).
- Aircraft design: [9] (25-dimensional domain; 500 simulations; response surfaces and Kriging), [10] (22-dimensional domain; minutes per simulation; response surfaces and Kriging), and [11] (31-dimensional domain; 20 minutes to several days per simulation; Kriging).
- Electric circuits: [12] (60-dimensional domain; 216 simulations; Kriging).

But how accurate are emulators? For Bayesian emulators, it is common to use the posterior distribution to measure uncertainty [13]. Another common technique for “validation” is to measure the error of the emulator on observations that were not used to train the emulator [14]. Both techniques require conditions which usually either cannot be verified or are known to be false. Posterior distributions depend on the choice of prior and likelihood. However, input variables generally represent fixed parameters, e.g., physical constants, and thus are not random.² Even truly stochastic parameters seldom have known distributions. Validation using held-out observations is relevant only if the error of the emulator at the held-out observations is representative of the error everywhere, or at least at the unobserved inputs that might matter. If the observations were independent and identically distributed (IID), representative samples could be obtained through random sampling. But values of f are clearly not IID: nearby inputs produce similar outputs, at least when the computer model has some regularity.³

Of course, without any conditions on f , there is no basis for extrapolating from the values of f at observed inputs to the values of f at unobserved inputs. Even though f may be “a black box”, it is often taken for granted that f is regular, in an unspecified sense. There are many ways regularity could be measured. If f is effectively a black box, there can be little scientific basis for measuring regularity in one way rather than another.

We work with what is known in numerical analysis as *the absolute condition number* and in function approximation as *the Lipschitz constant*. If f has Lipschitz constant K , then the difference between the output at two input points is not more than K times the

²By “random,” we mean stochastic. Bayesians use ‘probability’ to quantify uncertainties that are aleatoric (caused by randomness) and epistemic (caused by ignorance). Using probability to quantify one’s beliefs about a parameter does not make the parameter’s value random.

³One might try to correct for the correlation between nearby inputs, a well-known problem (for instance, [15]). But adjusting for correlation requires assuming that the observations follow a known distribution—the same problematic assumption that a fully Bayesian approach requires.

distance between those two points (measuring distance in some pre-specified metric). Similar results could be derived for other measures of regularity, but Lipschitz bounds are particularly amenable to analysis.

With perfect and complete knowledge of f , we could emulate it perfectly—by f itself. However, if the emulator \hat{f} is constrained to be computable from the observations, without relying on any other information about f , accuracy cannot be guaranteed. If it were known that the Lipschitz constant of f is K , that, together with the observations, would make possible a guarantee of some level of accuracy. All else equal, the larger K is, the more difficult it is to guarantee that an approximation of f is accurate.

The observations impose a lower bound on K (but no upper bound). We explore whether some approximation of f , computable from the observations alone, is guaranteed to be accurate throughout the domain of f —no matter what f is—provided f agrees with the observations and has a Lipschitz constant not greater than the observed lower bound on K . Viewed as a function of w in the domain of f , the minimax error of emulators over the set \mathcal{F} of functions that agree with the observations and have Lipschitz constant no greater than the lower bound is called the *potential error*.

The potential error is the “mini-minimax uncertainty” the title refers to. The first “mini” refers to the regularity condition: since f is essentially guaranteed to be less regular than the n observations require, the potential error is a lower bound on the minimax uncertainty for functions that are as regular as f . The second “mini” refers to emulators: this is the uncertainty for the best emulator. The “max” is over functions that agree with f at the observations and satisfy the optimistic regularity condition. That is, the potential error is the smallest that the maximum uncertainty could be, for the best emulator, for functions no less regular than f is required to be by the observations. The maximum potential error over the domain of f is the *maximum potential error*.

If K were known, finding the potential error would be a standard problem in information-based complexity [16, 17, 18]. However, K is unknown because f is only partially observed. We derive bounds on the potential error using a lower bound for K computed from the observed variation in f .

First, we derive a lower bound on the number of additional observations that could be necessary to learn f (section 3). We apply the bound to climate simulations from the Community Atmosphere Model, and find that the number of computer simulations needed to approximate f to within any useful accuracy could be astronomical.

Next, we give two lower bounds on the maximum potential error for approximating f from a fixed set of observations (section 4). The first bound is empirical. If this bound is large, then every emulator based on these observations is *potentially substantially incorrect*. The second bound, which is tighter, is a fraction of the unknown Lipschitz constant. That bound enables us to give conditions under which the constant emulator based on just one observation at the centroid has smaller maximum potential error than any emulator trained on the actual observations. When these conditions hold—and they hold for the climate simulations—every emulator is potentially substantially incorrect, even though there might be many observations, made at great expense.

We then extend the results for the maximum pointwise error over the domain of f to quantiles of the error across the domain, and to the mean of the error across the domain.

Finally, we consider additional information about f that would reduce the potential error, and the availability of such information in applications. We also allude to measures that might be more relevant than potential error in some scientific or engineering contexts.

2. Notation and problem formulation. The function f to be emulated is a fixed unknown real-valued function on $[0, 1]^p$, the p -dimensional unit cube. The space of real-valued continuous functions on $[0, 1]^p$ is $\mathcal{C}[0, 1]^p$. The Roman letters i, j, n, N, p , and q denote integers. Lowercase Greek letters denote real scalars, with the exception of μ , which denotes Lebesgue measure. Uppercase Roman letters such as X and D denote subsets of $[0, 1]^p$; X is a fixed finite subset of $[0, 1]^p$. Lowercase Roman letters from the end of the alphabet, such as v, w, x, y , and z , denote points in $[0, 1]^p$. The lowercase Roman letters e, f, g , and h denote real-valued functions on (subsets of) $[0, 1]^p$. The domain of the function g is $\text{dom}(g)$. The restriction of g to $D \subset \text{dom}(g)$ is denoted $g|_D$. The observations from which f is to be emulated are $f|_X$: we observe f on X . (We assume that the observations are noise-free; section 6 discusses adapting the method to noisy observations.) The function \hat{f} denotes a function on $[0, 1]^p$ that may be selected using the observations $f|_X$, but no other information about f ; it is an emulator of f . Let $\|h\|_\infty \equiv \sup_{w \in \text{dom}(h)} |h(w)|$, the infinity-norm of h . This paper concerns how large $\|\hat{f} - f\|_\infty$ could be, for the best possible choice of \hat{f} : the minimax pointwise error among emulators.

Let d be a metric on $\text{dom}(g)$. The (best) Lipschitz constant for g is

$$K(g) \equiv \sup \left\{ \frac{g(v) - g(w)}{d(v, w)} : v, w \in \text{dom}(g) \text{ and } v \neq w \right\}. \quad (2.1)$$

If $f \notin \mathcal{C}[0, 1]^p$, then $K(f) \equiv \infty$. Define

$$\mathcal{F}_\kappa(g) \equiv \{h \in \mathcal{C}[0, 1]^p : K(h) \leq \kappa \text{ and } h|_{\text{dom}(g)} = g\}.$$

Then $\mathcal{F}_\infty(f|_X)$ is the space of functions in $\mathcal{C}[0, 1]^p$ that fit the observations: they interpolate f on X .

Definition. The *potential error* of $\hat{f} \in \mathcal{C}[0, 1]^p$ over the set of functions \mathcal{F} is

$$\mathcal{E}(w; \hat{f}, \mathcal{F}) \equiv \sup \left\{ |\hat{f}(w) - g(w)| : g \in \mathcal{F} \right\}.$$

Definition. The *maximum potential error* of $\hat{f} \in \mathcal{C}[0, 1]^p$ over the set of functions \mathcal{F} is

$$\mathcal{E}(\hat{f}, \mathcal{F}) \equiv \sup_{w \in [0, 1]^p} \mathcal{E}(w; \hat{f}, \mathcal{F}) = \left\{ \|\hat{f} - g\|_\infty : g \in \mathcal{F} \right\}.$$

Maximum potential error is an instance of *worst-case error*, as defined in information-based complexity literature [16, 17, 18]. The uncertainty of the emulator \hat{f} can be quantified by its maximum potential error over the set $\mathcal{F}_\infty(f|_X)$ of continuous functions that agree with the observations, $\mathcal{E}(\hat{f}, \mathcal{F}_\infty(f|_X))$.

This measure of uncertainty presumes that $f \in \mathcal{C}[0, 1]^p$. If $f \notin \mathcal{C}[0, 1]^p$, \hat{f} could differ from f by even more. Our main results concern lower bounds on the uncertainty of the best possible emulator of f , under optimistic assumptions about the regularity of f . Why make

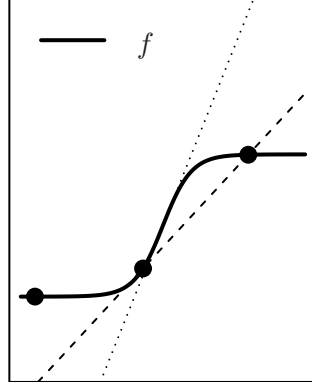


Figure 2.1. The dotted line is tangent to f where f attains its Lipschitz constant. Thus it has slope K . The dashed line is the steepest line that intersects any pair of observations. Thus it has slope \hat{K} . Clearly $\hat{K} \leq K$.

assumptions about the regularity of f ? The set X is not dense in $[0, 1]^p$, so for any $c > 0$, there exists some function $g \in \mathcal{F}_\infty(f|_X)$ satisfying $\|f - g\|_\infty > c$. Thus $\mathcal{E}(\hat{f}, \mathcal{F}_\infty(f|_X)) = \infty$: the maximum potential error is infinite unless f has more regularity than mere continuity.

Let $K \equiv K(f)$ and $\hat{K} \equiv K(f|_X)$. Because $X \subset [0, 1]^p$, $\hat{K} \leq K$. Figure 2.1 illustrates this inequality. In this and subsequent figures, $p = 1$ and the bold black dots represent $f|_X$, the observations of f at $x \in X$.

Henceforth, to simplify notation, we set

$$\mathcal{F}_\kappa \equiv \mathcal{F}_\kappa(f|_X)$$

and

$$\mathcal{E}_\kappa(\hat{f}) \equiv \mathcal{E}(\hat{f}, \mathcal{F}_\kappa).$$

Define the *radius* of $\mathcal{F} \subset \mathcal{C}[0, 1]^p$ to be

$$r(\mathcal{F}) \equiv \frac{1}{2} \sup \{ \|g - h\|_\infty : g, h \in \mathcal{F} \}.$$

By the triangle inequality, for any \hat{f} ,

$$\mathcal{E}_\kappa(\hat{f}) \geq r(\mathcal{F}_\kappa). \quad (2.2)$$

Equality holds for the emulator that “splits the difference”:

$$f_\kappa^*(w) \equiv \frac{1}{2} \left[\inf_{g \in \mathcal{F}_\kappa} g(w) + \sup_{g \in \mathcal{F}_\kappa} g(w) \right]$$

[16, 17]. That is, for all emulators \hat{f} that agree with f on X ,

$$\mathcal{E}_\kappa(\hat{f}) \geq \mathcal{E}_\kappa(f_\kappa^*) \equiv \mathcal{E}_\kappa^* :$$

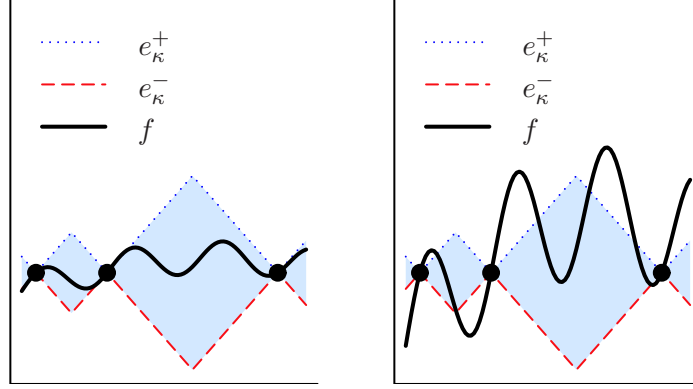


Figure 2.2. For both panels, $\hat{K} = 0$ and the optimal interpolant f_κ^* is constant. In the left panel $\kappa = K$. In the right panel $\kappa < K$. If $\kappa \geq K$ then $e_\kappa^- \leq f \leq e_\kappa^+$, and, equivalently, $f \in \mathcal{F}_\kappa$.

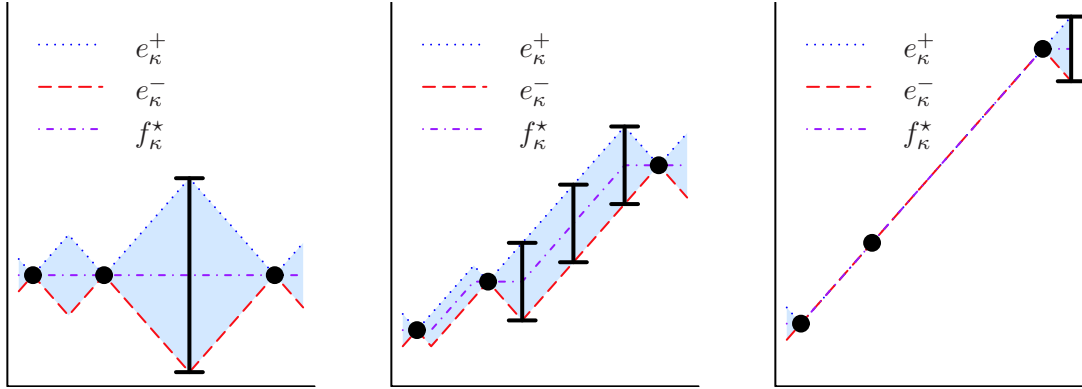


Figure 2.3. The length of the black error bars are twice the maximum potential error over \mathcal{F}_κ . The succession of panels shows that as the slope between observations approaches κ , $e^*(w)$ approaches 0 for points w between observations, and the maximum potential error over \mathcal{F}_κ decreases.

the emulator f_κ^* is a minimax (over $f \in \mathcal{F}_\kappa$) for infinity-norm error. The function f_κ^* depends implicitly on the observational design X through $\mathcal{F}_\kappa = \mathcal{F}_\kappa(f|_X)$. Because $\hat{K} \leq K$,

$$\mathcal{F}_{\hat{K}} \subset \mathcal{F}_K$$

and hence

$$\mathcal{E}_K^* \geq \mathcal{E}_{\hat{K}}^*. \quad (2.3)$$

That is, the minimax maximum potential error over the set of functions that agree with f on X and have Lipschitz constant K is at least as large as the minimax maximum potential error over the set of functions that agree with f on X and have Lipschitz constant equal to the empirical Lipschitz constant \hat{K} —the least regularity consistent with the observations $f|_X$.

Define

$$e_\kappa^+(w) \equiv e_{f,X,\kappa}^+(w) \equiv \min_{x \in X} [f(x) + \kappa d(x, w)],$$

$$e_{\kappa}^{-}(w) \equiv e_{f,X,\kappa}^{-}(w) \equiv \max_{x \in X} [f(x) - \kappa d(x, w)],$$

and

$$e_{\kappa}^{\star}(w) \equiv e_{f,X,\kappa}^{\star}(w) \equiv \frac{1}{2} [e_{f,X,\kappa}^{+}(w) - e_{f,X,\kappa}^{-}(w)].$$

The last of these, $e_{\kappa}^{\star}(w)$, measures minimax error at w : Consider the set of functions g that agree with the observations $f|_X$ and have Lipschitz constant no larger than κ . If $\kappa \leq K$, for all we know f could be any element of that set. Consider all possible emulators \hat{f} that also agree with the observations $f|_X$. The smallest (across emulators \hat{f}) maximum (across functions g) error at the point $w \in [0, 1]^p$ is $e_{\kappa}^{\star}(w)$. Figures 2.2 and 2.3 illustrate these definitions.

3. Bounds on the number of observations needed to approximate f well . To determine how many observations are required to approximate f , we must first decide how much error $\epsilon > 0$ to tolerate in the approximation.

Definition. If $\|\hat{f}|_A - g|_A\|_{\infty} \leq \epsilon$, then \hat{f} ϵ -approximates g on A . If $A = \text{dom}(g)$, then \hat{f} ϵ -approximates g .

Definition. If \mathcal{F} is a non-empty class of functions with common domain D , then \hat{f} ϵ -approximates \mathcal{F} on $A \subset D$ if $\forall g \in \mathcal{F}$, \hat{f} ϵ -approximates g on A . If $A = D$, then \hat{f} ϵ -approximates \mathcal{F} .

The emulator \hat{f} ϵ -approximates \mathcal{F} if and only if the maximum potential error of \hat{f} on \mathcal{F} does not exceed ϵ .

If ultimately we seek to ϵ -approximate \mathcal{F}_K , what is a reasonable value for ϵ ? Since \hat{K} is the observed variation of f on X , a useful value of ϵ would typically be much smaller than \hat{K} . (Otherwise, we might just as well take \hat{f} to be a constant; c.f. section 4.) For fixed f and ϵ , the number of observations needed to ensure that f_{κ}^{\star} ϵ -approximates \mathcal{F}_K depends on the experimental design X . The following definitions help to bound the number of observations needed, without restricting our analysis to a particular experimental design.

Definition. For fixed $\epsilon > 0$, and $Y \subset \text{dom}(f)$, Y is ϵ -adequate for f on A if f_K^{\star} ϵ -approximates $\mathcal{F}_K(f|_Y)$ on A . If $A = \text{dom}(f)$, then Y is ϵ -adequate for f .

Let $B(x, \delta)$ denote the open ball in \mathbb{R}^p centered at x with radius δ . Let

$$N_f \equiv \min\{\#Y : Y \text{ is } \epsilon\text{-adequate for } f\},$$

where $\#Y$ is the cardinality of Y .

Definition. The *minimum potential computational burden* is

$$M \equiv \max\{N_g : g \in \mathcal{F}_K\}.$$

This is also a minimax quantity: over all experimental designs Y , M is the smallest number of observations of f needed to guarantee that the maximum error of the best emulator based on those observations is not larger than ϵ . Minimum potential computational burden is an example of *minimum worst-case cost*, as defined in information-based complexity literature [17, 18].

3.1. An upper bound on N_f . For each $x \in X$, f_K^* ϵ -approximates $\mathcal{F}_K(f|_K)$ on (at least) $B(x, \epsilon/K)$. Thus, f_K^* ϵ -approximates \mathcal{F}_K on $\bigcup_{x \in X} B(x, \epsilon/K)$. Hence, the cardinality of any $Y \subset [0, 1]^p$ for which

$$V \equiv \left\{ B\left(x, \frac{\epsilon}{K}\right) : x \in Y \right\} \supset [0, 1]^p$$

is an upper bound on N_f . Because K is unknown, however, $\#V$ has limited utility as an upper bound on N_f in practice.

To generalize this formulation, let K^+ be a known upper bound on K . When d is such that $[0, 1]^p$ may be covered by N^+ balls of radius ϵ/K^+ , then $N_f \leq N^+$. While, in general, finding an optimal covering set is NP-hard, it is often straightforward to find good upper bounds, or even exact solutions, for some metrics. For example, in ℓ_∞ , $[0, 1]^p$ can be covered by $\left\lceil \frac{K^+}{2\epsilon} \right\rceil^p$ balls of radius ϵ/K^+ .

3.2. A lower bound on M . Depending on f and X , it can happen that f_K^* ϵ -approximates \mathcal{F}_K on regions of the domain not contained in $\bigcup_{x \in X} B(x, \epsilon/K)$. To see this, consider $p = 1$, $f(x) = x$, and $X = \{0, 1\}$. Then $K = \hat{K} = 1$. In this case, the observations $f|_X$ determine f exactly: the only function in \mathcal{F}_K is f . In this example, for a function g to agree with the observations requires it to attain the Lipschitz constant K everywhere. A function cannot agree with the observations and “run away” from f .

More generally, if f varies on X , then for a function g to agree with f at the observations requires g to vary too. That required variation “spends” some of g ’s Lipschitz constant, preventing g from running as far away from f as it could if f_X were constant. We now quantify this intuition to find lower bounds for M .

Define $\bar{\gamma} \equiv \arg \min_{\gamma \in \mathbb{R}} \sum_{x \in X} |f(x) - \gamma|^p$. Computing $\bar{\gamma}$ is straightforward because the objective function is univariate and convex.⁴ Let $X^+ \equiv \{x \in X : f(x) \geq \bar{\gamma}\}$ and let $X^- \equiv \{x \in X : f(x) < \bar{\gamma}\}$. Let

$$Q_+ \equiv \bigcup_{x \in X^+} \left\{ B\left(x, \frac{f(x) - \bar{\gamma}}{\hat{K}}\right) \cap [0, 1]^p \right\}$$

and

$$Q_- \equiv \bigcup_{x \in X^-} \left\{ B\left(x, \frac{\bar{\gamma} - f(x)}{\hat{K}}\right) \cap [0, 1]^p \right\}.$$

Then $Q_+ \cap Q_- = \emptyset$.⁵

⁴Alternatively, we could set $\bar{\gamma} \equiv \frac{1}{\#X} \sum_{x \in X} f(x)$, where $\#X$ is the size of X . The resulting lower bound may not be as tight.

⁵Fix $x^+ \in X^+$ and $x^- \in X^-$. Then $|f(x^+) - f(x^-)|/d(x^+, x^-) \leq \hat{K}$. Equivalently, $d(x^+, x^-) \geq |f(x^+) - f(x^-)|/\hat{K}$. Let $B^+ = B\left(x^+, [f(x^+) - \bar{\gamma}]/\hat{K}\right)$ and $B^- = B\left(x^-, [\bar{\gamma} - f(x^-)]/\hat{K}\right)$. Let a be the sum of the radii of B^+ and B^- . Then $a = (f(x^+) - \bar{\gamma})/\hat{K} + (\bar{\gamma} - f(x^-))/\hat{K} = (f(x^+) - f(x^-))/\hat{K}$, and $a \leq d(x^+, x^-)$. Therefore, $B^+ \cap B^- = \emptyset$. Because our selection of $x^+ \in X^+$ and $x^- \in X^-$ was arbitrary, $Q^+ \cap Q^- = \emptyset$.

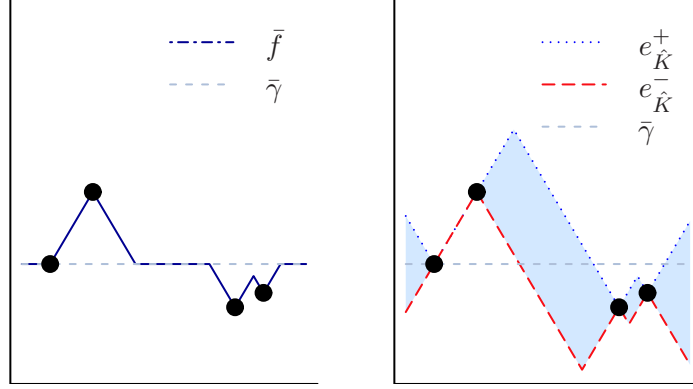


Figure 3.1. The function \bar{f} (shown in the left panel) is comprised of segments of $e_{\hat{K}}^+$, $e_{\hat{K}}^-$ and the constant function $\bar{\gamma}$ (all shown in the right panel). It is constant over roughly half of the domain. No function between $e_{\hat{K}}^-$ and $e_{\hat{K}}^+$ (inclusive) is constant over a larger fraction of the domain.

Define

$$\bar{f} : [0, 1]^p \rightarrow \mathbb{R}$$

$$w \mapsto \begin{cases} e_{\hat{K}}^-(w), & w \in Q_+ \\ e_{\hat{K}}^+(w), & w \in Q_- \\ \bar{\gamma}, & \text{otherwise.} \end{cases}$$

Figure 3.1 illustrates this definition. If we know $f|_X$, we know \bar{f} . By construction, $\bar{f} \in \mathcal{F}_{\hat{K}} \subset \mathcal{F}_K$.

Let $\bar{Q} \equiv [0, 1]^p \setminus (Q_+ \cup Q_-)$. Let μ be Lebesgue measure. By the union bound, because $\mu([0, 1]^p) = 1$,

$$\mu(\bar{Q}) \geq 1 - \sum_{x \in X} \mu\left(B\left(x, |f(x) - \bar{\gamma}|/\hat{K}\right)\right).$$

Let $C_2 \equiv \frac{\pi^{p/2}}{\Gamma(p/2+1)}$ and $C_\infty \equiv 2^p$, where Γ is the gamma function. Then, for $q \in \{2, \infty\}$,

$$\mu(\bar{Q}) \geq 1 - C_q \sum_{x \in X} \left(|f(x) - \bar{\gamma}|/\hat{K}\right)^p.$$

If there is some $x \in X$ for which the singleton set $\{x\}$ is ϵ -adequate for f on $A \subset \bar{Q}$, then $\mu(A) \leq \mu(B(0, \epsilon/\hat{K}))$. Hence, because $\bar{f} \in \mathcal{F}_K$,

$$\begin{aligned} M &\geq \left\lceil \frac{\mu(\bar{Q})}{\mu(B(0, \epsilon/\hat{K}))} \right\rceil \\ &\geq \left\lceil \epsilon^{-p} \left[\frac{\hat{K}^p}{C_q} - \sum_{x \in X} |f(x) - \bar{\gamma}|^p \right] \right\rceil. \end{aligned} \tag{3.1}$$

This lower bound can be extremely large when p is not small.

3.3. Application to climate modeling. Under the Uncertainty Quantification Strategic Initiative Laboratory-Directed Research and Development Project 10-SI-013, the Lawrence Livermore National Laboratory (LLNL) Institutional Science and Technology Office provided us with results from 1154 climate simulations using the Community Atmosphere Model (CAM). Each of the $p = 21$ parameters was scaled so that the interval $[0, 1]$ contained all values considered physically reasonable. The output of interest was a scalar, the simulated global average upwelling longwave flux (FLUT) approximately 50 years in the future. That is, the simulator could be thought of as a function f that maps $[0, 1]^p \rightarrow \mathbb{R}$. Observing f by running the simulator for a particular input was computationally expensive; each run took several days on a supercomputer. LLNL used several approaches to choose the points $X \subset [0, 1]^p$ at which to run simulations, including Latin hypercube, one-at-a-time, and random-walk multiple-one-at-a-time [5]. The 1154 simulations include all points selected by any of those approaches.

For these observations, $\bar{\gamma} = 232.77$; $\hat{K} = 14.20$ for $q = 2$. By (3.1),

$$M \geq \left\lceil \epsilon^{-21} \left[\frac{1.57 \times 10^{24}}{0.0038} - 6.81 \times 10^{24} \right] \right\rceil > \epsilon^{-21} \times 10^{26}.$$

For example, if ϵ is 1% of \hat{K} , then $M \geq 10^{43}$. Even if ϵ is 50% of \hat{K} , $M > 10^8$. For $q = \infty$, $\hat{K} = 34.68$; in that case

$$M \geq \left\lceil \epsilon^{-21} \left[\frac{2.19 \times 10^{32}}{2^{21}} - 6.81 \times 10^{25} \right] \right\rceil > \epsilon^{-21} \times 10^{25}.$$

These lower bounds on the minimum potential computational burden are extreme for a wide range of values of ϵ : there are functions that fit the observations and are no less regular than the observations require, but that cannot be approximated with useful precision from any tractable number of observations. The function \bar{f} , which is simple to construct, attains these lower bounds on potential computational burden.

4. Bounds on the maximum potential error for a fixed experimental design. The previous section gave lower bounds on the potential computational burden (cardinality of X) required to attain a desired maximum potential error ϵ . This section gives two lower bounds on the maximum potential error \mathcal{E}_K^* for a fixed experimental design X : an absolute bound and a bound expressed as a fraction of K . The bound as a fraction of K can yield a strong negative result: when a statistic—calculable from the observations—exceeds a calculable threshold, the maximum potential error is no less than the maximum potential error from a single observation at the centroid of the domain. That is, observing f for every point in X was (from the perspective of maximum potential error) wasteful: a single observation would have been better. In the CAM climate model example, maximum potential error is quite large, both in magnitude and as a fraction of K .

4.1. Lower bounds. For $g \in \mathcal{C}[0, 1]^p$, define

$$\inf g \equiv \inf \{g(w) : w \in [0, 1]^p\}$$

and

$$\sup g \equiv \sup \{g(w) : w \in [0, 1]^p\}.$$

Proposition 1. $\sup e_\kappa^* = \mathcal{E}_\kappa^*$.

The proof of this proposition and others is in A.

Because $\mathcal{F}_{\hat{K}} \subset \mathcal{F}_K$, for any emulator \hat{f} ,

$$\mathcal{E}_K(\hat{f}) \geq r(\mathcal{F}_{\hat{K}}). \quad (4.1)$$

Hence, by proposition 1, we have established we have established

Corollary 2. $\mathcal{E}_K(\hat{f}) \geq \sup e_{\hat{K}}^*$.

Corollary 2 is one of our principal results: it shows that $\sup e_{\hat{K}}^*$, a statistic calculable solely from the observations $f|_X$, is a lower bound on the maximum potential error for *any* emulator \hat{f} based on the observations $f|_X$.

Theorem 3 gives a stronger lower bound in terms of the unknown value of K .

Theorem 3. For any λ , if $\sup e_{\hat{K}}^* \geq \lambda \hat{K}$, then $\mathcal{E}_K(\hat{f}) \geq \lambda K$.

4.2. Maximum potential error for an emulator based on one observation. In sections 4.2 and 4.3 we work in ℓ_∞ : $d(v, w) = \|v - w\|_\infty$, which makes the calculations simpler and gives a particularly strong result.

Let $z \equiv (1/2, \dots, 1/2)$, the centroid of $[0, 1]^p$. Let $\hat{g} \in \mathcal{F}_\infty(f|_{\{z\}})$ be the constant function $\hat{g}(w) \equiv f(z)$, $\forall w \in [0, 1]^p$. The ℓ_∞ distance from z to any point on the boundary of $[0, 1]^p$ is $1/2$, so

$$\mathcal{E}_K(\hat{g}, \mathcal{F}_K(f|_{\{z\}})) = \frac{K}{2}.$$

That is, the maximum potential error of the emulator that is constant throughout $[0, 1]^p$ and equal to the value of f at the centroid of the cube is $K/2$. Let $W \subset [0, 1]^p$ be finite and $c \in \mathbb{R}$. Suppose $f|_W = c$. Let $\hat{h} \in \mathcal{F}_\infty(f|_W)$. By examining the corners of the domain, it follows that if $|W| < 2^p$,

$$\mathcal{E}_K(\hat{h}, \mathcal{F}_K(f|_W)) \geq \frac{K}{2}.$$

That is, if f is constant on W , any emulator based on fewer than 2^p observations of f will have at least $K/2$ maximum potential error. Making 2^p observations of f is intractable for CAM and many other applications. If f is nearly constant the situation may still be hopeless.

How do we know whether $f|_X$ is too close to constant to benefit from observing it more than once, but fewer than 2^p times?

Corollary 4. If $\sup e_{\hat{K}}^* \geq \hat{K}/2$, then

$$\mathcal{E}_K(\hat{f}) = \mathcal{E}_K(\hat{f}, \mathcal{F}_K(f|_X)) \geq \frac{K}{2} \geq \mathcal{E}_K(\hat{g}, \mathcal{F}_K(f|_{\{z\}})).$$

That is, if $\sup e_{\hat{K}}^* \geq \hat{K}/2$, no emulator based on observing $f|_X$ has smaller maximum potential error than the constant emulator based on a single observation— f is too nearly constant. Corollary 4 follows directly from theorem 3, taking $\lambda = \hat{K}/2$.

4.3. Application to climate modeling. We now return to the Community Atmosphere Model. Is the maximum potential error of the best emulator based on observing f at the 1154 points in X lower than the maximum potential error of the constant emulator based on one observation of f at the centroid of $[0, 1]^p$? We cannot simply compute these two maximum

potential errors, because K is unknown. But corollary 4 applies if we can determine whether $\sup e_{\hat{K}}^* \geq \hat{K}/2$. Recall that $\hat{K} = 34.68$ in ℓ_∞ for this dataset. Unfortunately, determining $\sup e_{\hat{K}}^*$ is difficult. In ℓ_∞ , if $f|_X$ is constant, it amounts to finding a maximal empty hypercube, a problem recently shown to be NP-hard in p [19]. It is generally no easier if f varies on X . Fortunately, it suffices to bound $\sup e_{\hat{K}}^*$. Since we are working in ℓ_∞ , we can bound $\sup e_{\hat{K}}^*$ above and below by considering just the corners of $[0, 1]^p$.

4.3.1. Computing an upper bound from non-adjacent corners in ℓ_∞ . Recall that throughout sections 4.2 and 4.3 $d(v, w) = \|v - w\|_\infty$.

Lemma 5. *Let $\mathbf{0} \equiv (0, \dots, 0)$ and $\mathbf{1} \equiv (1, \dots, 1)$. For $v, w \in [0, 1]^p$,*

$$d(v, w) \leq \tilde{d}(v) \equiv \max(d(v, \mathbf{0}), d(v, \mathbf{1})).$$

Proposition 6.

$$\sup e_{\hat{K}}^* \leq \frac{1}{2} \left\{ \min_{x \in X} [f(x) + \hat{K} \tilde{d}(x)] - \max_{x \in X} [f(x) - \hat{K} \tilde{d}(x)] \right\}.$$

Using this proposition, we calculate $\sup e_{\hat{K}}^* \leq 20.95$ for the CAM dataset.

4.3.2. Computing lower bounds from corners in ℓ_∞ . Clearly

$$\sup e_{\hat{K}}^* \geq \max \left\{ e_{\hat{K}}^*(w) : \forall w \in \{0, 1\}^p \right\}.$$

Perhaps surprisingly, this lower bound is essentially sharp for the CAM dataset. The domain, $[0, 1]^p$, contains 2^p corners $\{r_i\}_1^{2^p}$. Divide $[0, 1]^p$ into 2^p hypercubes $\{R_i\}_{i=1}^{2^p}$ with edge-length $1/2$, disjoint interiors, each containing a different corner of $[0, 1]^p$ (e.g., one such hypercube is $[0, 1/2]^p$). Then the R_i are disjoint ℓ_∞ -balls of radius $1/2$. Because X contains only 1154 points, the vast majority of the R_i do not contain any element of X . Because $e_{\hat{K}}^*$ tends to increase with distance from points in X , these unoccupied hypercubes are good regions to explore to find points maximizing $e_{\hat{K}}^*$. Within an unoccupied hypercube R_i , no point is farther in ℓ_∞ from any point in X than the corner r_i . So, the corners $\{r_i\}_1^{2^p}$ are good places to observe $e_{\hat{K}}^*$ for the purposes of establishing a tight lower bound on $\sup e_{\hat{K}}^*$.

For the CAM dataset, one corner r_j attains $e_{\hat{K}}^*(r_j) = 20.95$. So, $e_{\hat{K}}^*$ attains the upper bound established in the previous section, and we conclude that $\sup e_{\hat{K}}^* = 20.95$.

4.3.3. Implications for the Community Atmosphere Model. Because $\sup e_{\hat{K}}^* = 20.95 \geq 17.34 = \hat{K}/2$, theorem 3 says that $\mathcal{E}_K(\hat{f}) \geq K/2$ for any interpolation \hat{f} . In other words, by the discussion in section 4.2, our maximum potential error would have been no greater had we just observed f once, at z , and predicted $\hat{f}(w) = f(z)$ for all $w \in [0, 1]^p$.

In some sense, this result is not surprising: if we had fixed \hat{K} but replaced f with a constant function, and $\#X < 2^p$, then $\sup e_{\hat{K}}^* \geq \hat{K}/2$, with equality holding if and only if $z \in X$. By repeating the bounding procedures from the previous two sections with $\hat{K}/2 = 17.34$ fixed but f replaced with constant function c , we find $\sup e_{c, X, \hat{K}}^* = 26.95$. The increase in maximum potential error from 20.95 to 26.95 that results from replacing f with a constant shows that the observed variation in f reduces the maximum potential error considerably—although the maximum potential error remains quite large.

norm	95% lower confidence bound			
	lower quartile	median	upper quartile	average
Euclidean	1.454	1.596	1.731	1.595
supremum	0.649	0.717	0.782	0.715

Table 5.1

Error of the minimax emulator $f_{\hat{K}}^*$ of the CAM model based on the LLNL observations. Column 1: distance metric d used to define the Lipschitz constant. Columns 2–4: Binomial lower confidence bounds for quartiles of the pointwise error, obtained by inverting binomial tests. Column 5: 95% lower confidence bound for the integral of the pointwise error over the entire domain $[0, 1]^p$, based on inverting a z -test. Columns 2–5 are expressed as a fraction of $\hat{K}/2$ —half the empirical Lipschitz constant. Results are based on 10,000 uniform random samples from $[0, 1]^p$.

5. Extensions. The theory developed in this paper focuses on the maximum uncertainty over all $w \in [0, 1]^p$. We have argued that this maximum uncertainty is important in some applications. In other applications, whether the uncertainty is anywhere large might be less interesting than the fraction of the domain $[0, 1]^p$ on which the uncertainty is large.

It is possible to estimate the fraction of $[0, 1]^p$ for which e^* exceeds any fixed threshold $\epsilon \geq 0$ by sampling. By drawing values $w \in [0, 1]^p$ at random and evaluating e^* at each selected point, it is possible to construct a lower confidence bound for the fraction of the domain $[0, 1]^p$ for which the potential error exceeds any given threshold, and to construct confidence bounds for quantiles of the potential error.

Table 5.1 shows the results for the CAM simulations, based on 10,000 random samples from $[0, 1]^p$. Even the lower quartiles are a large fraction of \hat{K} . For instance, at confidence level 95%, the potential error under the sup-norm metric exceeds 71.7% of $\hat{K}/2$ on at least 50% of the domain.

6. Conclusions. We find a lower bound on the minimum (over emulators) maximum (over regular functions that agree with the observations) error of emulators of a function f based on n observations. This “mini-minimax” uncertainty is optimistic because it assumes that f is no less regular (i.e., not “rougher”) than the n observations require f to be. It measures the *potential error* of the best emulator of f at the point w in the domain of f : there are functions g and h that are no less regular than f , that agree with f at the n observations, and that differ at the point w by twice this potential error. No emulator can approximate both functions with less error than the potential error at w . The observations and the regularity condition do not rule out g or h : the true f *could* be either.

In some problems, *every* emulator based on any tractable number of observations of f has large maximum potential error (and the potential error is large over much of the domain), even if f is no less regular than it is observed to be. That is, there are functions g and h that agree perfectly with the observations, are no less regular than the observations require them to be, and yet differ by a large amount at some point in the domain of f . When that occurs, all emulators of f are *potentially substantially incorrect*.

We give sufficient conditions under which all emulators are potentially substantially incorrect. The conditions depend only on the observed values of f ; they can be computed from the same observations used to train an emulator, at a cost that typically is small compared

with the cost of generating those observations. The conditions are sufficient but not necessary, because f could be less regular than any finite set of observations reveals it to be. It is not possible to give necessary conditions that depend only on the observed values of f (a priori bounds on the regularity of f would be needed).

Our conditions seem likely to hold for many HEB applications of societal interest, such as climate modeling: the curse of (*High*) dimensionality makes it necessary to observe f at many points, unless it is known a priori that f is extremely regular; the *Expense* makes it intractable to observe f at that many points; and the *Black-box* nature of f means we do not typically know a priori how regular f is. Indeed, we show quantitatively that these conditions hold for a common measure of regularity and a large climate modeling dataset.

When the maximum potential error in approximating f everywhere by a constant—the value of f at the center of the domain—is no larger than the maximum potential error in approximating f from any tractable number of observations, emulators may not be useful. In that circumstance, no emulator can reliably model f as a function of its input $w \in [0, 1]^p$.

Common techniques for validating emulators do not reveal whether emulators are potentially substantially incorrect, either because they make strong assumptions about f that are based neither on the observations nor on real knowledge of the properties of f (e.g., Bayesian approaches), or because they estimate the average error with respect to some measure, rather than the potential error (e.g., validation using held-out observations). Our results reflect the long-standing tension between worst-case analysis and average-case analysis,⁶ between Bayesian and minimax measures of uncertainty, and between simple, high-bias models and complex models that are harder to estimate and interpret. The complexity of many HEB models may be on the far side of the tradeoff, necessitating the use of emulators that are destined to be highly uncertain because computational constraints preclude adequate sampling.

Reducing the potential error of emulators in HEB problems requires either more information about f (knowledge, not merely assumptions⁷), or changing the measure of uncertainty—changing the scientific question. Both tactics are application-specific: the underlying science dictates the conditions that actually hold for f and the senses in which it is useful to approximate f .

Finally, it is not clear that emulators help address the most interesting questions. Approximating f pointwise is not typically the ultimate goal; in many problems, most properties of f are nuisance parameters. It could well be that the important questions about f can be answered more directly, without estimating f as a whole. For example, for global optimisation—finding maxima or minima—a form of adaptive sampling known as multi-start methods yield good results [21]. It may be possible to develop an adaptive approach to finding level sets—another common use for emulators—too [22].

There may not always be a remedy, however. Some research questions simply cannot be answered through simulation at present. Employing complex emulators and massive compu-

⁶However, as section 5 shows, the average potential error for the CAM model is also quite large.

⁷Common additional conditions include the following: parameters have only low order interactions; the second derivative has an upper bound; the third derivative has a limited number of knots; the integral of the squared derivative of the model is bounded [20]. There are problems in which conditions like these may reflect actual knowledge about f . However, such conditions tend to be difficult to verify: simulation is perhaps most valuable when the underlying equations are not amenable to mathematical analysis.

tational resources can distract us from this reality. With greater scientific understanding of the underlying function, it may become possible to reduce uncertainty in HEB simulations to useful levels. Until then, simpler models—with caveats about what they omit—may be a better basis for scientific inference.

REFERENCES

- [1] J Sacks, WJ Welch, TJ Mitchell, and HP Wynn. Design and Analysis of Computer Experiments. *Statistical Science*, 1989.
- [2] EN Ben-Ari and DM Steinberg. Modeling data from computer experiments: An empirical comparison of Kriging with MARS and projection pursuit regression. *Quality Engineering*, 2007.
- [3] RG Ghanem, A Doostan, and J Red-Horse. A probabilistic construction of model validation. *Computer Methods in Applied Mechanics and Engineering*, 2008.
- [4] S Shan and GG Wang. Survey of modeling and optimization strategies to solve high-dimensional design problems with computationally-expensive black-box functions. *Structural and Multidisciplinary Optimization*, 2009.
- [5] C Covey, S Brandon, PT Bremer, D Domyancis, X Garaizar, G Johannesson, R Klein, SA Klein, DD Lucas, J Tannahill, and Y Zhang. A new ensemble of perturbed-input-parameter simulations by the Community Atmosphere Model. Technical report, Lawrence Livermore National Laboratory, 2011.
- [6] D Aspenberg, J Jergeus, and L Nilsson. Robust optimization of front members in a full frontal car impact. *Engineering Optimization*, 2012.
- [7] M Holena, D Linke, and U Rodemerck. Generator approach to evolutionary optimization of catalysts and its integration with surrogate modeling. *Catalysis Today*, 2011.
- [8] JA Shorter, PC Ip, and HA Rabitz. An efficient chemical kinetics solver using high dimensional model representation. *The Journal of Physical Chemistry A*, 1999.
- [9] A Srivastava, K Hacker, K Lewis, and TW Simpson. A method for using legacy data for metamodel-based design of large-scale systems. *Structural and Multidisciplinary Optimization*, 2004.
- [10] PN Koch, TW Simpson, and JK Allen. Statistical approximations for multidisciplinary design optimization: the problem of size. *Journal of Aircraft*, 1999.
- [11] AJ Booker, JE Dennis, PD Frank, DB Serafini, V Torczon, and Trosset MW. A rigorous framework for optimization of expensive functions by surrogates. *Optimization*, 1999.
- [12] RA Bates, RJ Buck, E Riccomagno, and HP Wynn. Experimental design and observation for large systems. *Journal of the Royal Statistical Society, Series B*, 1996.
- [13] C Tebaldi and RL Smith. Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *Journal of Climate*, 2005.
- [14] K Fang, R Li, and A Sudjianto. *Design and modeling for computer experiments*. 2006.
- [15] LS Bastos and A O'Hagan. Diagnostics for gaussian process emulators. *Technometrics*, 2009.
- [16] EW Packel. Do linear problems have linear optimal algorithms? *SIAM Review*, 1988.
- [17] J Traub and H Woźniakowski. *A general theory of optimal algorithms*. 1980.
- [18] JF Traub, GW Wasilkowski, and H Woźniakowski. *Information-based complexity*. 1988.
- [19] J Backer and JM Keil. The mono- and bichromatic empty rectangle and square problems in all dimensions. In *LATIN 2010: Theoretical Informatics*, 2010.
- [20] M Lamboni, B Iooss, AL Popelin, and F Gamboa. Derivative-based global sensitivity measures: general links with Sobol'indices and numerical tests. *arXiv preprint*, 2012.
- [21] FJ Hickernell. A simple multistart algorithm for global optimization. *OR Transactions*, 1997.
- [22] JMJ Huttunen and PB Stark. Cheap contouring of costly functions: the Pilot Approximation Trajectory algorithm. *Computational Science & Discovery*, 2012.

Appendix A. Proofs.

To start, we relate the maximum potential error to an intersection of intervals in the range of f . For real χ and ρ , define the interval (one-dimensional ball)

$$I(\chi, \rho) \equiv \begin{cases} [\chi - \rho, \chi + \rho], & \rho \geq 0 \\ \emptyset, & \text{otherwise.} \end{cases}$$

If I is an interval, $\mu(I)$ denotes its length; for instance, $\mu(I(\chi, \rho)) = \max(0, 2\rho)$.

For $g \in \mathcal{C}[0, 1]^p$ and $\kappa \in \mathbb{R}$,

$$\begin{aligned} e_{g,X,\kappa}^*(w) &= \frac{1}{2} \left[e_{g,X,\kappa}^+(w) - e_{g,X,\kappa}^-(w) \right] \\ &= \frac{1}{2} \left\{ \min_{x \in X} \{g(x) + \kappa d(x, w)\} - \max_{x \in X} \{g(x) - \kappa d(x, w)\} \right\} \\ &= \frac{1}{2} \mu \left(\left[\max_{x \in X} \{g(x) - \kappa d(x, w)\}, \min_{x \in X} \{g(x) + \kappa d(x, w)\} \right] \right) \\ &= \frac{1}{2} \mu \left(\bigcap_{x \in X} I(g(x), \kappa d(x, w)) \right). \end{aligned} \tag{A.1}$$

Lemma 7. Fix $\alpha \in [0, 1]$, $\rho_1, \dots, \rho_n \in [0, \infty)$ and $\chi_1, \dots, \chi_n \in \mathbb{R}$. Let $I_1 \equiv \bigcap_{i=1}^n I(\chi_i, \rho_i)$ and $I_\alpha \equiv \bigcap_{i=1}^n I(\chi_i, \alpha \rho_i)$. Then $\alpha \mu(I_1) \geq \mu(I_\alpha)$.

Proof. Because the intersection of intervals is itself an interval, there exist χ_0 and ρ_0 satisfying

$$I_\alpha = I(\chi_0, \rho_0).$$

Fix $i \in 1, \dots, n$. Then

$$I(\chi_0, \rho_0) \subset I(\chi_i, \alpha \rho_i).$$

It follows that

$$\chi_0 - \rho_0 \geq \chi_i - \alpha \rho_i.$$

Then

$$\alpha \left(\rho_i - \frac{\rho_0}{\alpha} \right) \geq \chi_i - \chi_0.$$

Because $\alpha \leq 1$ and $\rho_i \geq 0$,

$$\rho_i - \frac{\rho_0}{\alpha} \geq \chi_i - \chi_0.$$

Finally,

$$\chi_0 - \frac{\rho_0}{\alpha} \geq \chi_i - \rho_i.$$

By symmetric reasoning we also have that

$$\chi_0 + \frac{\rho_0}{\alpha} \leq \chi_i + \rho_i.$$

Therefore,

$$I\left(\chi_0, \frac{\rho_0}{\alpha}\right) \subset I(\chi_i, \rho_i).$$

Because i was arbitrary,

$$I\left(\chi_0, \frac{\rho_0}{\alpha}\right) \subset I_1.$$

Hence,

$$\mu(I_1) \geq \mu\left(I\left(\chi_0, \frac{\rho_0}{\alpha}\right)\right) = \frac{2\rho_0}{\alpha} = \frac{\mu(I_\alpha)}{\alpha}.$$

□

Lemma 7 is used in the proof of Theorem 3, below.

Proposition 1. $\sup e_\kappa^* = \mathcal{E}_\kappa^*$.

Proof.

Step 1: e_κ^+ and e_κ^- are continuous.

For fixed $x \in X$, if $g \in \mathcal{C}[0, 1]^p$, $g(x) \pm \kappa d(x, w)$ is continuous in w . Because the minimum or maximum of a finite set of continuous functions is also continuous, $\min_{x \in X} [g(x) + \kappa d(x, w)]$ and $\max_{x \in X} [g(x) - \kappa d(x, w)]$ are continuous in w too. Thus e_κ^+ and e_κ^- are continuous functions on $[0, 1]^p$.

Step 2: e_κ^+ and e_κ^- agree with f on X .

Suppose g has Lipschitz constant κ and g agrees with f on X . Then for any $x \in X$ and $w \in [0, 1]^p$,

$$f(x) - \kappa d(w, x) \leq g(w) \leq f(x) + \kappa d(w, x).$$

Hence,

$$\max_{x \in X} [f(x) - \kappa d(x, w)] \leq g(w) \leq \min_{x \in X} [f(x) + \kappa d(x, w)].$$

If $w \in X$,

$$\min_{x \in X} [f(x) + \kappa d(x, w)] = \max_{x \in X} [f(x) - \kappa d(x, w)] = f(w).$$

Thus, e_κ^+ and $e_\kappa^- \in \mathcal{F}_\kappa$ agree with f on X .

Step 3: e_κ^+ and e_κ^- have Lipschitz constant κ .

For $v, w \in [0, 1]^p$, $\exists x, y \in X$ satisfying

$$e_\kappa^+(v) = f(x) + \kappa d(x, v) \text{ and } e_\kappa^+(w) = f(y) + \kappa d(y, w).$$

Suppose without loss of generality that $e^+(v) \geq e^+(w)$. By construction, $e_\kappa^+(v) \leq f(y) + \kappa d(y, v)$. Hence

$$\begin{aligned} 0 &\leq e_\kappa^+(v) - e_\kappa^+(w) \leq f(y) + \kappa d(y, v) - e_\kappa^+(w) \\ &= f(y) + \kappa d(y, v) - f(y) - \kappa d(y, w) \\ &\leq \kappa(d(y, v) - d(y, w)) \\ &\leq \kappa d(v, w), \end{aligned} \tag{A.2}$$

by the triangle inequality. Hence e_κ^+ has Lipschitz constant κ . An analogous argument shows that e_κ^- also has Lipschitz constant κ . Combining this with steps 1 and 2 shows that e_κ^+ and e_κ^- are in \mathcal{F}_κ .

Step 4: e_κ^- is the pointwise minimum of \mathcal{F}_κ and e_κ^+ is the pointwise maximum of \mathcal{F}_κ . Suppose to the contrary that there exists $w \in [0, 1]^p$, $x \in X$, and $g \in \mathcal{F}_\kappa$ for which

$$g(w) > f(x) + \kappa d(x, w).$$

Recall that $g \in \mathcal{F}_\kappa$ implies that $g(x) = f(x) \forall x \in X$. Hence

$$g(w) - g(x) > f(x) + \kappa d(x, w) - f(x) = \kappa d(x, w).$$

That is, g has a Lipschitz constant greater than κ , a contradiction. Hence, $e_\kappa^+(w) = \sup\{g(w) : g \in \mathcal{F}_\kappa\}$ for all $w \in [0, 1]^p$. The same argument, *mutatis mutandi*, shows that $e_\kappa^-(w) = \inf\{g(w) : g \in \mathcal{F}_\kappa\}$ for all $w \in [0, 1]^p$.

Step 5: $\sup e_\kappa^* = \mathcal{E}_\kappa^*$.

Now

$$\begin{aligned} \sup_{w \in [0, 1]^p} e_\kappa^*(w) &= \sup_{w \in [0, 1]^p} \frac{1}{2} [e_\kappa^+(w) - e_\kappa^-(w)] \\ &= \frac{1}{2} \sup_{w \in [0, 1]^p} \left\{ \sup_{g \in \mathcal{F}_\kappa} g(w) - \inf_{h \in \mathcal{F}_\kappa} h(w) \right\} \\ &= \frac{1}{2} \sup_{w \in [0, 1]^p} \left\{ \sup_{g, h \in \mathcal{F}_\kappa} |g(w) - h(w)| \right\} \\ &= \frac{1}{2} \sup_{g, h \in \mathcal{F}_\kappa} \sup_{w \in [0, 1]^p} |g(w) - h(w)| \\ &= \frac{1}{2} \sup \{ \|g - h\|_\infty : g, h \in \mathcal{F}_\kappa \} \\ &= \mathcal{E}_\kappa^*. \end{aligned}$$

□

Theorem 3. For any λ , if $\sup e_{\hat{K}}^* \geq \lambda \hat{K}$, then $\mathcal{E}_K(\hat{f}) \geq \lambda K$.

Proof. Let $w^* \equiv \arg \max_w e_{\hat{K}}^*(w)$. Then

$$\begin{aligned} \mathcal{E}_K(\hat{f}) &\geq \mathcal{E}_K(f_K^*) \\ &= \sup e_K^* \end{aligned} \tag{A.3}$$

$$\begin{aligned} &\geq e_K^*(w^*) \\ &\geq \frac{K}{\hat{K}} \cdot e_{\hat{K}}^*(w^*) \end{aligned} \tag{A.4}$$

$$\begin{aligned} &\geq \frac{K}{\hat{K}} \cdot \lambda \hat{K} \\ &= \lambda K. \end{aligned} \tag{A.5}$$

(A.3) is a consequence of proposition 1. (A.5) follows from (A.4) by hypothesis. (A.4) is a consequence of lemma 7: Let $\alpha = \hat{K}/K \leq 1$. For, $i = 1, \dots, \#X$, let $\rho_i = f(x_i)$ and $\chi_i = Kd(x, w)$. Then, by (A.1), $\mu(I_1)/2 = e_K$ and $\mu(I_\alpha)/2 = e_{\hat{K}}$. □

Lemma 5. Let $\mathbf{0} \equiv (0, \dots, 0)$ and $\mathbf{1} \equiv (1, \dots, 1)$. For $v, w \in [0, 1]^p$,

$$d(v, w) \leq \tilde{d}(v) \equiv \max(d(v, \mathbf{0}), d(v, \mathbf{1})).$$

Proof. Let $w_{(i)}$ denote the i^{th} component of w . Then

$$\begin{aligned} d(v, w) &= \max_{i \in \{1, \dots, p\}} |v_{(i)} - w_{(i)}| \\ &\leq \max_{i \in \{1, \dots, p\}} \max_{\delta \in \{0, 1\}} |v_{(i)} - \delta| \\ &= \max_{i \in \{1, \dots, p\}} \max_{y \in \{\mathbf{0}, \mathbf{1}\}} |v_{(i)} - y_{(i)}| \\ &= \max_{y \in \{\mathbf{0}, \mathbf{1}\}} \max_{i \in \{1, \dots, p\}} |v_{(i)} - y_{(i)}| \\ &= \max_{y \in \{\mathbf{0}, \mathbf{1}\}} d(v, y) \\ &= \max(d(v, \mathbf{0}), d(v, \mathbf{1})). \end{aligned}$$

□

Proposition 6.

$$\sup e_{\hat{K}}^* \leq \frac{1}{2} \left\{ \min_{x \in X} [f(x) + \hat{K} \tilde{d}(x)] - \max_{x \in X} [f(x) - \hat{K} \tilde{d}(x)] \right\}.$$

Proof. Fix $w \in [0, 1]^p$. Then,

$$e_{\hat{K}}^*(w) = \frac{1}{2} \mu \left(\bigcap_{x \in X} I(f(x), \hat{K} d(x, w)) \right) \tag{A.6}$$

$$\leq \frac{1}{2} \mu \left(\bigcap_{x \in X} I(f(x), \hat{K} \tilde{d}(x)) \right) \tag{A.7}$$

$$= \frac{1}{2} \left\{ \min_{x \in X} [f(x) + \hat{K} \tilde{d}(x)] - \max_{x \in X} [f(x) - \hat{K} \tilde{d}(x)] \right\}$$

where (A.6) follows from (A.1), and (A.7) follows from lemma 5. Because the right-hand side of this inequality does not depend on w , the proposition follows by taking suprema. □